Convex Optimization

Lijun Zhang

Nanjing University, China

December 24, 2017

Outline

Outline

Mathematical Optimization

- Optimization Problem [Boyd and Vandenberghe, 2004]

$$\begin{aligned} \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, n \end{aligned}$$

  - $\mathbf{x} = [x_1, \ldots, x_d]^\top \in \mathbb{R}^d$: optimization variable
  - $f_0 : \mathbb{R}^d \mapsto \mathbb{R}$: objective function
  - $f_i : \mathbb{R}^d \mapsto \mathbb{R}, i = 1, \ldots, n$: constraint functions

- Optimal Solution

$$\{\mathbf{x}_* | f_0(\mathbf{x}_*) \leq f_0(\mathbf{x}), \forall \mathbf{x} \in \Omega\}$$

where $\Omega = \{\mathbf{x} | f_i(\mathbf{x}) \leq 0, i = 1, \ldots, n\}$

LAMDA Learning And Mining from DatA

Least-squares

The Problem

$$\min_{\mathbf{x}} \quad f_0(\mathbf{x}) = \sum_{i=1}^{n} \left(\mathbf{a}_i^\top \mathbf{x} - b_i\right)^2 = \|A\mathbf{x} - \mathbf{b}\|^2$$

- $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^\top \in \mathbb{R}^{n \times d}$
- $\mathbf{b} = [b_1, \ldots, b_n]^\top \in \mathbb{R}^n$

Solving Least-squares

- Analytical solution: $\mathbf{x}_* = (A^\top A)^{-1} A^\top \mathbf{b}$
- Reliable and efficient algorithms and software
- Computation time proportional to $d^2 n$
- A mature technology

Linear Programming

The Problem

$$\begin{aligned} \min_{\mathbf{x}} \quad & \mathbf{c}^\top \mathbf{x} \\ \text{s.t.} \quad & \mathbf{a}_i^\top \mathbf{x} \leq b_i, \quad i = 1, \ldots, n \end{aligned}$$

- $\mathbf{x} = [x_1, \ldots, x_d]^\top$: optimization variable
- $\mathbf{c}, \mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^d$ and $b_1, \ldots, b_n \in \mathbb{R}$: parameters

Solving Linear Programs

- No analytical formula for solution
- Reliable and efficient algorithms and software
- Computation time proportional to $d^2 n$ if $n \geq d$
- A mature technology

Convex Optimization Problem

- The Problem

$$\begin{aligned} \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, n \end{aligned}$$

  - Objective and constraint functions are convex:

$$f_i(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f_i(\mathbf{x}) + (1-\theta) f_i(\mathbf{y})$$

    if $0 \leq \theta \leq 1$

- Solving Convex Optimization Problems
  - No analytical solution
  - Reliable and efficient algorithms
  - Almost a technology

Outline

Convex Set

- Line Segment between $\mathbf{x}_1$ and $\mathbf{x}_2$

$$\mathbf{x} = \theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2, \ 0 \leq \theta \leq 1$$

Convex Set

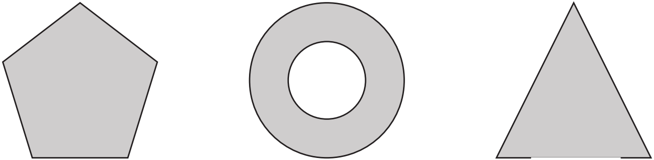
- Line Segment between $\mathbf{x}_1$ and $\mathbf{x}_2$

$$\mathbf{x} = \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2, \; 0 \leq \theta \leq 1$$

Convex Set

A set $\mathcal{C}$ is convex if the line segment between any two points in $\mathcal{C}$ lies in $\mathcal{C}$.

$$\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}, 0 \leq \theta \leq 1 \Rightarrow \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathcal{C}$$



Convex Functions

Convex Functions

A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is convex if $\operatorname{dom} f$ is a convex set and if for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$, and θ with $0 \leq \theta \leq 1$, we have

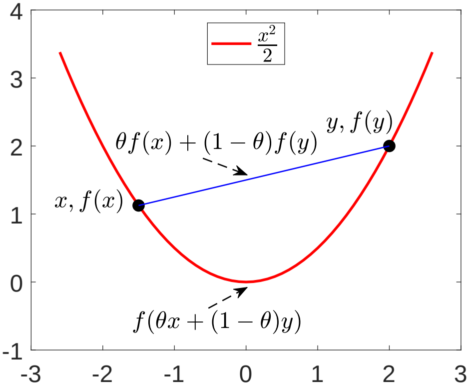
$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y})$$

Convex Functions

Convex Functions

A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is convex if $\operatorname{dom} f$ is a convex set and if for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$, and θ with $0 \le \theta \le 1$, we have

$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \le \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$$



Convex Functions

Convex Functions

A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is convex if $\operatorname{dom} f$ is a convex set and if for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$, and θ with $0 \leq \theta \leq 1$, we have

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y})$$

Concave Functions

A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is concave if $\operatorname{dom} f$ is a convex set and if for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$, and θ with $0 \leq \theta \leq 1$, we have

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \geq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y})$$

- f is concave if $-f$ is convex

Examples

- Examples on $\mathbb{R}$
  - Affine: $ax + b$ on $\mathbb{R}$, for any $a, b \in \mathbb{R}$
  - Exponential: e^{ax} on $\mathbb{R}$, for any $a \in \mathbb{R}$
  - Powers: x^a on $\mathbb{R}_{++}$, for $a \geq 1$ or $a \leq 0$
  - Powers of absolute value: $|x|^p$ on $\mathbb{R}$, for $p \geq 1$
  - Negative entropy: $x \log x$ on $\mathbb{R}_{++}$
- Examples on $\mathbb{R}^d$
  - Affine function: $\mathbf{a}^\top \mathbf{x} + b$, for any $\mathbf{a} \in \mathbb{R}^d$, $b \in \mathbb{R}$
  - Norm: $\|\mathbf{x}\|_p = (\sum_{i=1}^d |x_i|^p)^{1/p}$, for any $p \geq 1$

First-order Condition

- The Gradient of $f(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_d} \right]^\top \in \mathbb{R}^d$$

First-order Condition

- The Gradient of $f(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$

$$\nabla f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_d}\right]^\top \in \mathbb{R}^d$$

First-order Condition

Suppose f is differentiable. Then f is convex if and only if $\operatorname{dom} f$ is convex and

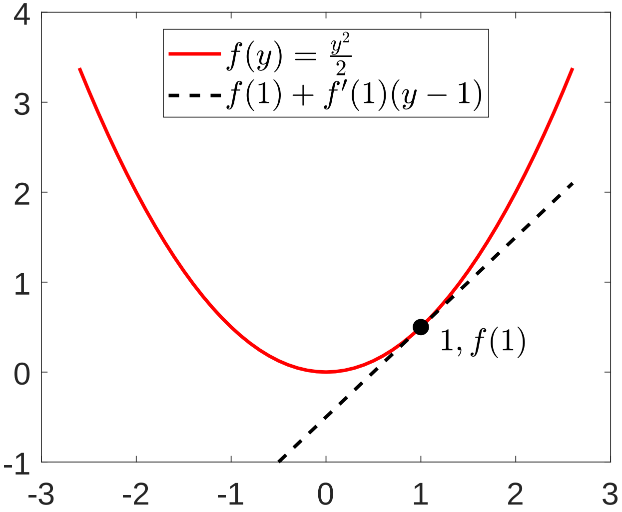
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

holds for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$.

- First-order approximation of f is a global underestimator

First-order Condition

- An Exmaple



Second-order Condition

- The Hessian of $f(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$

$$\nabla^2 f(\mathbf{x}) \in \mathbb{R}^{d \times d} \text{ with } \nabla^2 f(\mathbf{x})_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$$

Second-order Condition

- The Hessian of $f(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$

$$\nabla^2 f(\mathbf{x}) \in \mathbb{R}^{d \times d} \text{ with } \nabla^2 f(\mathbf{x})_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$$

Second-order Condition

Suppose f is twice differentiable. Then f is convex if and only if $\operatorname{dom} f$ is convex and its Hessian is positive semidefinite:

$$\nabla^2 f(\mathbf{x}) \succeq 0$$

holds for all $\mathbf{x} \operatorname{dom} f$.

LAMDA Learning And Mining from DatA

Second-order Condition

- The Hessian of $f(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$

$$\nabla^2 f(\mathbf{x}) \in \mathbb{R}^{d \times d} \text{ with } \nabla^2 f(\mathbf{x})_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}$$

Second-order Condition

Suppose f is twice differentiable. Then f is convex if and only if $\operatorname{dom} f$ is convex and its Hessian is positive semidefinite:

$$\nabla^2 f(\mathbf{x}) \succeq 0$$

holds for all $\mathbf{x} \operatorname{dom} f$.

- Quadratic functions

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$$

where $A \succeq 0$

Operations that Preserve Convexity

Nonnegative Weighted Sum

- If $f_1, \ldots, f_n$ are convex and $w_1, \ldots, w_n \geq 0$, then

$$f = w_1 f_1 + w_2 f_2 + \cdots, w_n f_n \text{ is convex.}$$

Composition with Affine Function

- If f is convex, then

$$g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b}) \text{ is convex.}$$

Pointwise Maximum and Supremum

- If $f_1, \ldots, f_n$ are convex, then

$$f(x) = \max\{f_1(x), \ldots, f_n(x)\} \text{ is convex.}$$

- Hinge loss: $\ell(\mathbf{w}) = \max(0, 1 - y\mathbf{x}^\top \mathbf{w})$
- If for each $y \in \mathcal{A}$, $f(x, y)$ is convex in x, then the function

$$g(x) = \sup_{y \in \mathcal{A}} f(x, y) \text{ is convex.}$$

Jensen's inequality

- Basic Inequality
  - If f is convex, then for $0 \leq \theta \leq 1$,
$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$$

- Extensions
  - If f is convex, then
$$f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)]$$
where X is a random variable
  - Examples:
$$(\mathrm{E}[X])^2 \leq \mathrm{E}\left[X^2\right]$$
$$\mathrm{E}\left[\sqrt{X}\right] \leq \sqrt{\mathrm{E}[X]}$$

LAMDA Learning And Mining from DatA

Convex Optimization Problem

- Standard Form

$$\begin{aligned} \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, n \\ & \mathbf{a}_i^\top \mathbf{x} = b_i, \quad i = 1, \ldots, p \end{aligned}$$

  - $f_0, f_1, \ldots, f_n$ are convex
  - Equality constraints are affine

Convex Optimization Problem

- Standard Form

$$\begin{aligned} \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, n \\ & \mathbf{a}_i^\top \mathbf{x} = b_i, \quad i = 1, \ldots, p \end{aligned}$$

  - $f_0, f_1, \ldots, f_n$ are convex
  - Equality constraints are affine

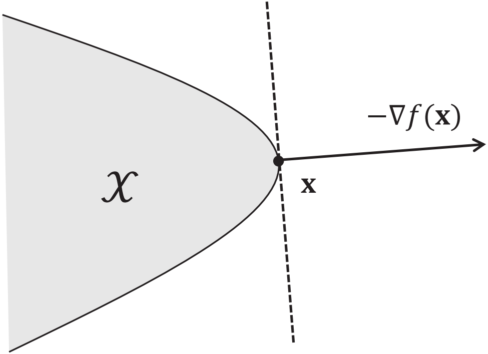
- Local and Global Optima

  Any locally optimal point of a convex problem is (globally) optimal.

Optimality Criterion for Differentiable f_0

- Optimality Criterion
  - $\mathbf{x}$ is optimal if and only if it is feasible and

$$\nabla f_0(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) \geq 0 \text{ for all feasible } \mathbf{y}$$



  - Unconstrained: $\mathbf{x}$ is optimal if and only if $\nabla f_0(\mathbf{x}) = 0$

The Conjugate Function

Conjugate Function

The conjugate of function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is defined as

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \operatorname{dom} f} \left(\mathbf{y}^\top \mathbf{x} - f(\mathbf{x}) \right)$$

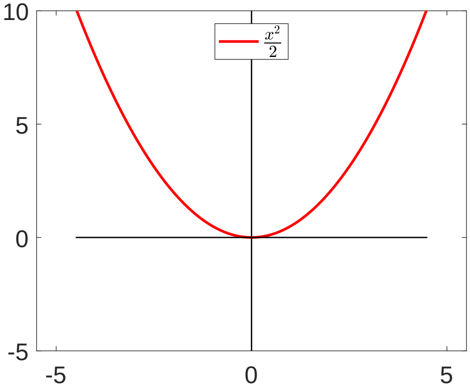
- $\operatorname{dom} f^*$
$$\left\{ \mathbf{y} \middle| \sup_{\mathbf{x} \in \operatorname{dom} f} \left(\mathbf{y}^\top \mathbf{x} - f(\mathbf{x}) \right) < \infty \right\}$$
- f^* is convex
- Suppose f is differentiable
$$f^*(\mathbf{y}) = \nabla f(\mathbf{x})^\top \mathbf{x} - f(\mathbf{x}) = -\left[f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (0 - \mathbf{x}) \right]$$
where $\mathbf{y} = \nabla f(\mathbf{x})$

The Conjugate Function

Conjugate Function

The conjugate of function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is defined as

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \operatorname{dom} f} \left(\mathbf{y}^\top \mathbf{x} - f(\mathbf{x}) \right)$$

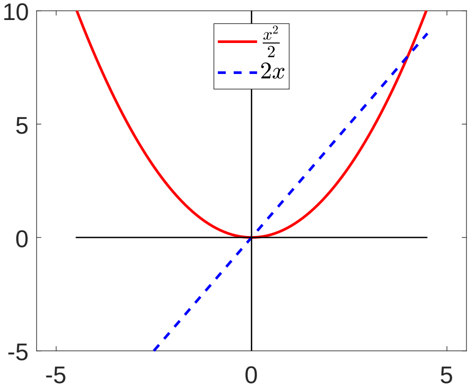


The Conjugate Function

Conjugate Function

The conjugate of function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is defined as

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \operatorname{dom} f} \left(\mathbf{y}^\top \mathbf{x} - f(\mathbf{x}) \right)$$

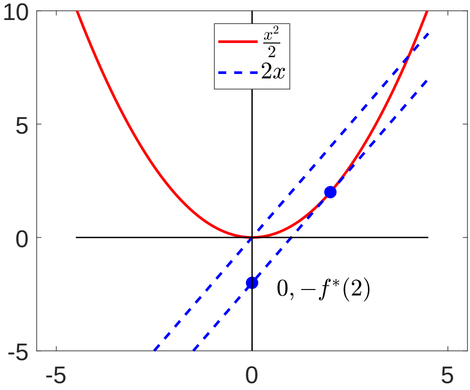


The Conjugate Function

Conjugate Function

The conjugate of function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is defined as

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \operatorname{dom} f} \left(\mathbf{y}^\top \mathbf{x} - f(\mathbf{x}) \right)$$



Outline

The Lagrangian

- The Optimization Problem (not necessarily convex)

$$\begin{aligned} \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{s. t.} \quad & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, n \\ & h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, p \end{aligned}$$

  - Variable $\mathbf{x} \in \mathbb{R}^d$, domain $\mathcal{D}$, optimal value p_*

- The Lagrangian $L : \mathbb{R}^d \times \mathbb{R}^n \times \mathbb{R}^p \mapsto \mathbb{R}$

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x})$$

  - λ_i is Lagrange multiplier associated with $f_i(\mathbf{x}) \leq 0$
  - ν_i is Lagrange multiplier associated with $h_i(\mathbf{x}) = 0$

Lagrange Dual Function

- Lagrange dual function $g : \mathbb{R}^n \times \mathbb{R}^p \mapsto \mathbb{R}$

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$
$$= \inf_{\mathbf{x} \in \mathcal{D}} \left\{ f_0(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x}) \right\}$$

  - g is concave

- Lower Bound Property
  - If $\boldsymbol{\lambda} \succeq 0$, then $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \le p_*$.

$$f_0(\tilde{\mathbf{x}}) \ge L(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \ge \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = g(\boldsymbol{\lambda}, \boldsymbol{\nu})$$

The Dual Problem

Lagrange Dual Problem

$$\begin{aligned} \max_{\boldsymbol{\lambda}, \boldsymbol{\nu}} \quad & g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{s. t.} \quad & \boldsymbol{\lambda} \succeq 0 \end{aligned}$$

- The best lower bound on p_*
- A convex optimization problem
- Optimal value d_*

The Dual Problem

- Lagrange Dual Problem

$$\begin{aligned} \max_{\boldsymbol{\lambda},\boldsymbol{\nu}} \quad & g(\boldsymbol{\lambda},\boldsymbol{\nu}) \\ \text{s. t.} \quad & \boldsymbol{\lambda} \succeq 0 \end{aligned}$$

  - The best lower bound on p_*
  - A convex optimization problem
  - Optimal value d_*

- Weak Duality

$$d_* \leq p_*$$

- Strong Duality

$$d_* = p_*$$

  - Doesnot hold in general, but usually holds for convex problems

Complementary Slackness

- Under Strong Duality

$$f_0(\mathbf{x}_*) = g(\boldsymbol{\lambda}_*, \boldsymbol{\nu}_*) = \inf_{\mathbf{x}\in\mathcal{D}} \left\{ f_0(\mathbf{x}) + \sum_{i=1}^{n} \lambda_{*i} f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_{*i} h_i(\mathbf{x}) \right\}$$

$$\leq f_0(\mathbf{x}_*) + \sum_{i=1}^{n} \lambda_{*i} f_i(\mathbf{x}_*) + \sum_{i=1}^{p} \nu_{*i} h_i(\mathbf{x}_*) \leq f_0(\mathbf{x}_*)$$

  - $\mathbf{x}_*$ is primal optimal and $(\boldsymbol{\lambda}_*, \boldsymbol{\nu}_*)$ is dual optimal

Complementary Slackness

- Under Strong Duality

$$
\begin{aligned}
f_0(\mathbf{x}_*) = g(\boldsymbol{\lambda}_*, \boldsymbol{\nu}_*) = \inf_{\mathbf{x}\in\mathcal{D}} &\left\{ f_0(\mathbf{x}) + \sum_{i=1}^{n} \lambda_{*i} f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_{*i} h_i(\mathbf{x}) \right\} \\
\leq & f_0(\mathbf{x}_*) + \sum_{i=1}^{n} \lambda_{*i} f_i(\mathbf{x}_*) + \sum_{i=1}^{p} \nu_{*i} h_i(\mathbf{x}_*) \leq f_0(\mathbf{x}_*)
\end{aligned}
$$

  - $\mathbf{x}_*$ is primal optimal and $(\boldsymbol{\lambda}_*, \boldsymbol{\nu}_*)$ is dual optimal

- Complementary Slackness

$$
\lambda_{*i} f_i(\mathbf{x}_*) = 0 \Rightarrow \begin{cases} \lambda_{*i} > 0 \Rightarrow f_i(\mathbf{x}_*) = 0 \\ f_i(\mathbf{x}_*) < 0 \Rightarrow \lambda_{*i} = 0 \end{cases}, \quad i = 1, \ldots, n
$$

- Optimality of $\mathbf{x}_*$

$$
\mathbf{x}_* = \operatorname*{arg\,min}_{\mathbf{x}\in\mathcal{D}} \left\{ f_0(\mathbf{x}) + \sum_{i=1}^{n} \lambda_{*i} f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_{*i} h_i(\mathbf{x}) \right\}
$$

Karush-Kuhn-Tucker (KKT) Conditions

1. Primal constraints

$$f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, n$$
$$h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, p$$

2. Dual constraints

$$\boldsymbol{\lambda} \succeq 0$$

3. Complementary Slackness

$$\lambda_{*i} f_i(\mathbf{x}_*) = 0, \quad i = 1, \ldots, n$$

4. Gradient of Lagrangian with respect to **x** vanishes

$$\nabla f_0(\mathbf{x}) + \sum_{i=1}^{n} \lambda_i \nabla f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i \nabla h_i(\mathbf{x}) = 0$$

Implications of KKT Conditions

General Problems

For any optimization problem for which strong duality obtains, any pair of primal and dual optimal points must satisfy the KKT conditions.

Implications of KKT Conditions

General Problems

For any optimization problem for which strong duality obtains, any pair of primal and dual optimal points must satisfy the KKT conditions.

Convex Optimizations

For any convex optimization problem, any points that satisfy the KKT conditions are primal and dual optimal, and have zero duality gap.

Implications of KKT Conditions

General Problems

For any optimization problem for which strong duality obtains, any pair of primal and dual optimal points must satisfy the KKT conditions.

Convex Optimizations

For any convex optimization problem, any points that satisfy the KKT conditions are primal and dual optimal, and have zero duality gap.

Convex Optimizations

For any convex optimization problem for which strong duality obtains, $\mathbf{x}_*$ is optimal if and only if there exist $(\boldsymbol{\lambda}_*, \boldsymbol{\nu}_*)$ that satisfy KKT conditions.

Support Vector Machines I

- The Optimization Problem

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \quad \sum_{i=1}^{n} \max\left(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

Support Vector Machines I

- The Optimization Problem

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \quad \sum_{i=1}^{n} \max\left(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

- Redefine the Hinge Loss

$$\ell(x) = \max(0, 1 - x)$$

- Its Conjugate Function

$$\ell^*(y) = \sup_x (yx - \ell(x)) = \begin{cases} y, & -1 \leq y \leq 0 \\ \infty, & \text{otherwise} \end{cases}$$

Support Vector Machines I

- The Optimization Problem

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \quad \sum_{i=1}^{n} \max\left(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

- Redefine the Hinge Loss

$$\ell(x) = \max(0, 1 - x)$$

- Its Conjugate Function

$$\ell^*(y) = \sup_x (yx - \ell(x)) = \begin{cases} y, & -1 \leq y \leq 0 \\ \infty, & \text{otherwise} \end{cases}$$

- A General Problem

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \quad \sum_{i=1}^{n} \ell\left(y_i(\mathbf{w}^\top \mathbf{x}_i + b)\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

Support Vector Machines II

An Equivalent Form

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}, \mathbf{u}\in\mathbb{R}^n} \quad & \sum_{i=1}^n \ell(u_i) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \\ \text{s. t.} \quad & u_i = y_i(\mathbf{w}^\top \mathbf{x}_i + b),\ i = 1\ldots,n \end{aligned}$$

Support Vector Machines II

- An Equivalent Form

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}, \mathbf{u}\in\mathbb{R}^n} \quad & \sum_{i=1}^{n} \ell(u_i) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \\ \text{s. t.} \quad & u_i = y_i(\mathbf{w}^\top \mathbf{x}_i + b),\ i = 1\ldots, n \end{aligned}$$

- The Lagrangian

$$\mathcal{L}(\mathbf{w}, b, \mathbf{u}, \boldsymbol{\nu}) = \sum_{i=1}^{n} \ell(u_i) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \nu_i \left(u_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\right)$$

Support Vector Machines II

- An Equivalent Form

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}, \mathbf{u}\in\mathbb{R}^n} \quad & \sum_{i=1}^{n} \ell(u_i) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \\ \text{s. t.} \quad & u_i = y_i(\mathbf{w}^\top \mathbf{x}_i + b),\ i = 1\ldots,n \end{aligned}$$

- The Lagrangian

$$\mathcal{L}(\mathbf{w}, b, \mathbf{u}, \boldsymbol{\nu}) = \sum_{i=1}^{n} \ell(u_i) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \nu_i \left(u_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b) \right)$$

- The Lagrange Dual Function

$$\begin{aligned} g(\boldsymbol{\nu}) &= \inf_{\mathbf{w}, b, \mathbf{u}} \mathcal{L}(\mathbf{w}, b, \mathbf{u}, \boldsymbol{\nu}) \\ &= \inf_{\mathbf{w}, b, \mathbf{u}} \sum_{i=1}^{n} \ell(u_i) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \nu_i \left(u_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b) \right) \end{aligned}$$

LAMDA Learning And Mining from DatA

Support Vector Machines III

- The Lagrange Dual Function

$$g(\boldsymbol{\nu})$$

$$= \inf_{\mathbf{w},b,\mathbf{u}} \sum_{i=1}^{n} (\ell(u_i) + \nu_i u_i) + \left(\frac{\lambda}{2} \|\mathbf{w}\|_2^2 - \mathbf{w}^\top \sum_{i=1}^{n} \nu_i y_i \mathbf{x}_i \right) - b \sum_{i=1}^{n} \nu_i y_i$$

Support Vector Machines III

- The Lagrange Dual Function

$$g(\boldsymbol{\nu})$$

$$= \inf_{\mathbf{w},b,\mathbf{u}} \sum_{i=1}^{n} (\ell(u_i) + \nu_i u_i)) + \left(\frac{\lambda}{2} \|\mathbf{w}\|_2^2 - \mathbf{w}^\top \sum_{i=1}^{n} \nu_i y_i \mathbf{x}_i \right) - b \sum_{i=1}^{n} \nu_i y_i$$

- Minimizing $\mathbf{w}, b, \mathbf{u}$

$$\inf_{u_i} (\ell(u_i) + \nu_i u_i)) = -\sup_{u_i} (-\nu_i u_i - \ell(u_i))$$

$$= -\ell^*(-\nu_i) = \nu_i, \text{ if } 0 \leq \nu_i \leq 1$$

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \mathbf{u}, \boldsymbol{\nu}) = \lambda \mathbf{w} - \sum_{i=1}^{n} \nu_i y_i \mathbf{x}_i = 0 \Rightarrow \mathbf{w} = \frac{1}{\lambda} \sum_{i=1}^{n} \nu_i y_i \mathbf{x}_i$$

$$\nabla_b \mathcal{L}(\mathbf{w}, b, \mathbf{u}, \boldsymbol{\nu}) = -\sum_{i=1}^{n} \nu_i y_i = 0$$

Support Vector Machines IV

- The Lagrange Dual Function

$$g(\boldsymbol{\nu}) = \sum_{i=1}^{n} \nu_i - \frac{1}{2\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} \nu_i \nu_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

Support Vector Machines IV

- The Lagrange Dual Function

$$g(\boldsymbol{\nu}) = \sum_{i=1}^{n} \nu_i - \frac{1}{2\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} \nu_i \nu_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

- The Dual Problem

$$\begin{aligned} \max_{\boldsymbol{\nu} \in \mathbb{R}^n} \quad & \sum_{i=1}^{n} \nu_i - \frac{1}{2\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} \nu_i \nu_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j \\ \text{s.t.} \quad & 0 \leq \nu_i \leq 1, \ i = 1 \ldots, n \\ \text{s.t.} \quad & \sum_{i=1}^{n} \nu_i y_i = 0 \end{aligned}$$

Support Vector Machines V

- KKT Conditions
  - $(\mathbf{w}_*, b_*, \mathbf{u}_*)$ and $\boldsymbol{\nu}_*$ are primal and dual solutions

$$u_{*i} = y_i(\mathbf{w}_*^\top \mathbf{x}_i + b_*)$$

$$0 \leq \nu_{*i} \leq 1$$

$$\sum_{i=1}^{n} \nu_{*i} y_i = 0$$

$$\mathbf{w}_* = \frac{1}{\lambda} \sum_{i=1}^{n} \nu_{*i} y_i \mathbf{x}_i$$

$$u_{*i} = \operatorname*{argmin}_{u_i} \left(\ell(u_i) + \nu_{*i} u_i\right) = 1 \text{ if } 0 < \nu_{*i} < 1$$

Outline

Performance Measure

- The Optimization Problem

$$\min_{\mathbf{w}\in\mathcal{W}} \quad f(\mathbf{w})$$

  - $f(\cdot)$ is a convex function
  - $\mathcal{W}$ is a convex domain

Performance Measure

- The Optimization Problem

$$\min_{\mathbf{w}\in\mathcal{W}} \quad f(\mathbf{w})$$

  - $f(\cdot)$ is a convex function
  - $\mathcal{W}$ is a convex domain

- Convergence Rate
  - After T iterations, the gap between objectives

$$f(\mathbf{w}_T) - f(\mathbf{w}_*) = O\left(\frac{1}{\sqrt{T}}\right), \ O\left(\frac{1}{T}\right), \ O\left(\frac{1}{T^2}\right), O\left(\frac{1}{\alpha^T}\right)$$

- Iteration Complexity
  - To ensure $f(\mathbf{w}_T) - f(\mathbf{w}_*) \leq \epsilon$, the order of T

$$T = \Omega\left(\frac{1}{\epsilon^2}\right), \ \Omega\left(\frac{1}{\epsilon}\right), \ \Omega\left(\frac{1}{\sqrt{\epsilon}}\right), \Omega\left(\log\frac{1}{\epsilon}\right)$$

LAMDA Learning And Mining from DatA

Analytical Properties

- Lipschitz Continuous

$$\|\nabla f(\mathbf{x})\| \leq G, \text{ or } \|\partial f(\mathbf{x})\| \leq G$$

- Strongly Convex

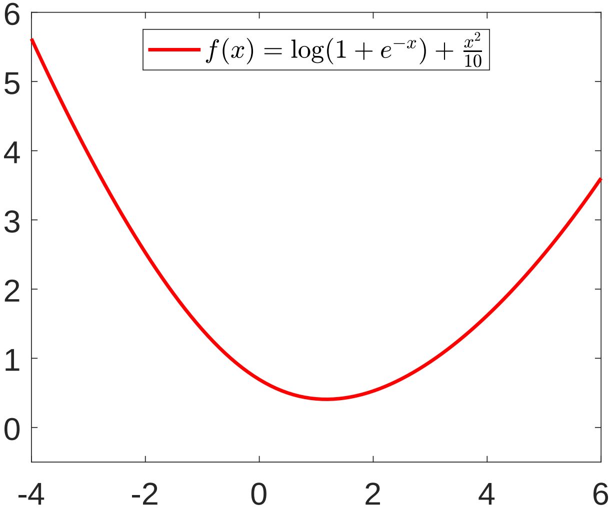
$$\nabla^2 f(\mathbf{x}) \succeq \mu I$$
$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu \|\mathbf{x} - \mathbf{y}\|_2^2$$
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2$$
$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}) - \theta(1-\theta)\frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2$$

- Smooth

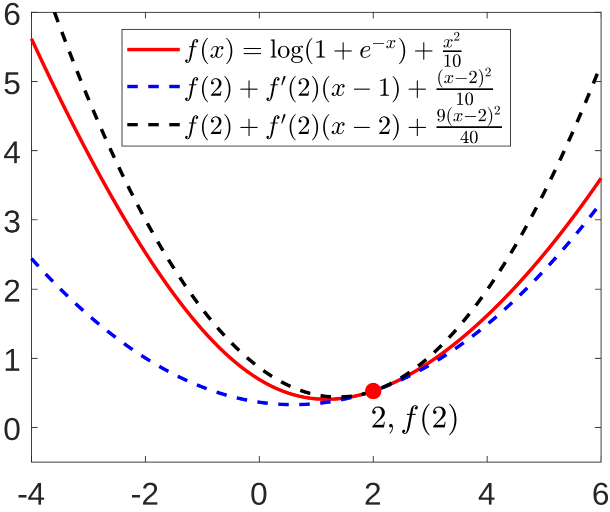
$$\nabla^2 f(\mathbf{x}) \preceq L I$$
$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq L \|\mathbf{x} - \mathbf{y}\|_2^2$$
$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

LAMDA
Learning And Mining from DatA

An Example



An Example



$f(x) = \log(1+e^{-x}) + \frac{x^2}{10}$
$f(2) + f'(2)(x-1) + \frac{(x-2)^2}{10}$
$f(2) + f'(2)(x-2) + \frac{9(x-2)^2}{40}$
$2, f(2)$
6
5
4
3
2
1
0
-4
-2
0
2
4
6

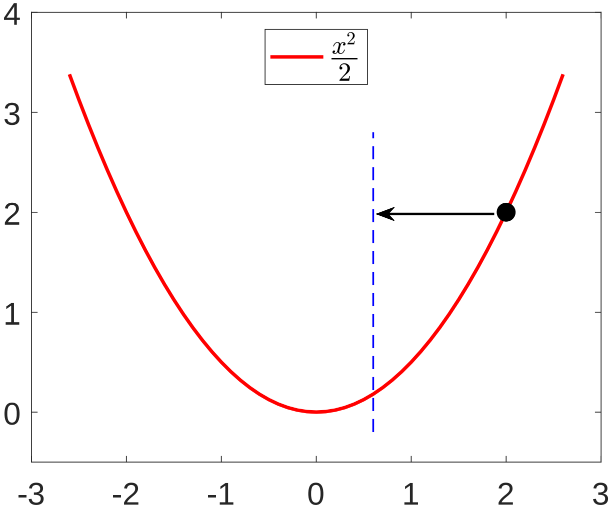
Existing Results

- Convergence Rate

Lipschitz Continuous	Strongly Convex	Smooth	Smooth Strongly Convex
GD $O\left(\frac{1}{\sqrt{T}}\right)$	EGD/SGD$_\alpha$ $O\left(\frac{1}{T}\right)$	AGD $O\left(\frac{1}{T^2}\right)$	GD/AGD $O\left(\frac{1}{\alpha^T}\right)$

- GD—Gradient Descent [Nesterov, 2004]
- EGD—Epoch Gradient Descent [Hazan and Kale, 2011]
- SGD$_\alpha$—SGD with α-suffix Averaging [Rakhlin et al., 2012]
- AGD—Nesterov's Accelerated Gradient Descent [Nesterov, 2005, Nesterov, 2007, Tseng, 2008]

Gradient Descent



Gradient Descent with Projection

- The Problem

$$\min_{\mathbf{w}\in\mathcal{W}} f(\mathbf{w})$$

for $t = 1, \ldots, T$ **do**

$$\mathbf{w}'_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)$$

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}'_{t+1})$$

end for

return $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$

- The Projection Operator

$$\Pi_{\mathcal{W}}(\mathbf{y}) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{W}} \|\mathbf{x} - \mathbf{y}\|_2$$

The Convergence Rate I

For any $\mathbf{w} \in \mathcal{W}$, we have

$$
\begin{aligned}
& f(\mathbf{w}_t) - f(\mathbf{w}) \\
\leq & \langle \nabla f(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w} \rangle \\
= & \frac{1}{\eta_t} \langle \mathbf{w}_t - \mathbf{w}'_{t+1}, \mathbf{w}_t - \mathbf{w} \rangle \\
= & \frac{1}{2\eta_t} \left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}'_{t+1} - \mathbf{w}\|_2^2 + \|\mathbf{w}_t - \mathbf{w}'_{t+1}\|_2^2 \right) \\
= & \frac{1}{2\eta_t} \left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}'_{t+1} - \mathbf{w}\|_2^2 \right) + \frac{\eta_t}{2} \|\nabla f(\mathbf{w}_t)\|_2^2 \\
\leq & \frac{1}{2\eta_t} \left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 \right) + \frac{\eta_t}{2} \|\nabla f(\mathbf{w}_t)\|_2^2
\end{aligned}
$$

To simplify the above inequality, we assume

$$\eta_t = \eta, \|\nabla f(\mathbf{w})\|_2 \leq G, \ \forall \mathbf{w} \in \mathcal{W}, \text{ and } \|\mathbf{x} - \mathbf{y}\|_2 \leq D, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{W}$$

The Convergence Rate II

Then, we have

$$f(\mathbf{w}_t) - f(\mathbf{w}) \leq \frac{1}{2\eta}\left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2\right) + \frac{\eta}{2}G^2$$

By adding the inequalities of all iterations, we have

$$\begin{aligned}
&\sum_{t=1}^{T} f(\mathbf{w}_t) - Tf(\mathbf{w}) \\
\leq & \frac{1}{2\eta}\left(\|\mathbf{w}_1 - \mathbf{w}\|_2^2 - \|\mathbf{w}_{T+1} - \mathbf{w}\|_2^2\right) + \frac{\eta T}{2}G^2 \\
\leq & \frac{1}{2\eta}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \frac{\eta T}{2}G^2 \\
\leq & \frac{1}{2\eta}D^2 + \frac{\eta T}{2}G^2 = GD\sqrt{T}
\end{aligned}$$

where we set

$$\eta = \frac{D}{G\sqrt{T}}$$

The Convergence Rate III

Then, we have

$$f(\bar{\mathbf{w}}_T) - f(\mathbf{w}) = f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t\right) - f(\mathbf{w})$$

$$\leq \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}_t) - f(\mathbf{w}) \leq \frac{1}{T} GD\sqrt{T} = \frac{GD}{\sqrt{T}}$$

The Convergence Rate III

Then, we have

$$f(\bar{\mathbf{w}}_T) - f(\mathbf{w}) = f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t\right) - f(\mathbf{w})$$

$$\leq \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}_t) - f(\mathbf{w}) \leq \frac{1}{T} GD\sqrt{T} = \frac{GD}{\sqrt{T}}$$

- Limitations
  - The step size $\eta = \frac{D}{G\sqrt{T}}$ depends on T
  - $f(\bar{\mathbf{w}}_T)$ doesnot decrease monotonically

Gradients or Subgradients

- The Logit Loss

$$\ell_i(\mathbf{w}) = \log\left(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right)$$

$$\nabla \ell_i(\mathbf{w}) = \frac{1}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla\left(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})\right)$$

$$= \frac{\exp(-y_i \mathbf{x}_i^\top \mathbf{w})}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} \nabla(-y_i \mathbf{x}_i^\top \mathbf{w}) = \frac{\exp(-y_i \mathbf{x}_i^\top \mathbf{w})}{1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})} - y_i \mathbf{x}_i$$

- The Hinge Loss

$$\ell_i(\mathbf{w}) = \max(0, 1 - y_i \mathbf{x}_i^\top \mathbf{w})$$

$$\partial \ell_i(\mathbf{w}) = \begin{cases} -y_i \mathbf{x}_i, & y_i \mathbf{x}_i^\top \mathbf{w} < 1 \\ 0, & y_i \mathbf{x}_i^\top \mathbf{w} > 1 \\ \{-\alpha y_i \mathbf{x}_i : \alpha \in [0,1]\}, & y_i \mathbf{x}_i^\top \mathbf{w} = 1 \end{cases}$$

Homework

Analyze the smoothness of the following function

$$f(x) = \log\left(1 + \exp(-x)\right) + \frac{\lambda}{2}x^2$$

$\Pi_{\mathcal{W}}(\cdot)$ is a nonexpanding operator

$$\|\Pi_{\mathcal{W}}(\mathbf{x}) - \Pi_{\mathcal{W}}(\mathbf{y})\|_2 \leq \|\mathbf{x} - \mathbf{y}\|_2, \forall \mathbf{x}, \mathbf{y}$$

The analysis of varied step size

$$\eta_t = O(1/\sqrt{t})$$

Reference I

Boyd, S. and Vandenberghe, L. (2004).
Convex Optimization.
Cambridge University Press.

Hazan, E. and Kale, S. (2011).
Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization.
In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436.

Nesterov, Y. (2004).
Introductory lectures on convex optimization: a basic course, volume 87 of *Applied optimization*.
Kluwer Academic Publishers.

Nesterov, Y. (2005).
Smooth minimization of non-smooth functions.
Mathematical Programming, 103(1):127–152.

Nesterov, Y. (2007).
Gradient methods for minimizing composite objective function.
Core discussion papers.

Reference II

Rakhlin, A., Shamir, O., and Sridharan, K. (2012).
Making gradient descent optimal for strongly convex stochastic optimization.
In *Proceedings of the 29th International Conference on Machine Learning*, pages 449–456.

Tseng, P. (2008).
On acclerated proximal gradient methods for convex-concave optimization.
Technical report, University of Washington.